JNCI 18-1296R1 – Accepted for publication

Article

Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists

Alejandro Rodriguez-Ruiz, MSc[1]; Kristina Lång, MD PhD[2]; Albert Gubern-Merida, PhD[3]; Mireille Broeders, PhD[4,5]; Gisella Gennaro, PhD[6]; Paola Clauser, MD[7]; Thomas H. Helbich, MD[7]; Margarita Chevalier, PhD[8]; Tao Tan, PhD[3]; Thomas Mertelmeier, PhD[9]; Matthew G. Wallis, MD[10]; Ingvar Andersson, MD PhD[11]; Sophia Zackrisson, MD PhD[12]; Ritse M. Mann, MD PhD[1]; Ioannis Sechopoulos, PhD[1,5]

[1]Department of Radiology and Nuclear Medicine, Radboud University Medical Center, PO Box 9101, 6500 HB Nijmegen, The Netherlands

[2]Institute for Biomedical Engineering, ETH Zürich, Gloriastrasse 35, 8092, Zürich, Switzerland

[3]ScreenPoint Medical BV, Stationplein 26, 6512 AB, Nijmegen, The Netherlands

[4]Department for Health Evidence, Radboud University Medical Center, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

[5]Dutch Expert Centre for Screening (LRCB), Wijchenseweg 101, 6538 SW, Nijmegen, The Netherlands

[6]Veneto Institute of Oncology (IOV)–IRCCS, via Gattamelata 64, 35128 Padua, Italy

[7]Department of Biomedical Imaging and Image-Guided Therapy, Division of Molecular and Gender Imaging, Medical University of Vienna, Waehringer Guertel 18-20, 1090 Vienna, Austria

[8]Medical Physics Group, Radiology Department, Faculty of Medicine, Universidad Complutense de Madrid, Pza. Ramón y Cajal s/n, 28040 Madrid, Spain

[9]Siemens Healthcare GmbH, Diagnostic Imaging, X-Ray Products, Technology & Concepts, Siemensstr. 3, 91301 Forchheim, Germany

[10]Cambridge Breast Unit and NIHR Biomedical Research Unit, Box 97, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Hills Road, CB2 0QQ, Cambridge, United Kingdom

[11]Unilabs Breast Center, Skåne University Hospital, Jan Waldenströms gata 22, SE-20502 Malmö, Sweden

[12]Diagnostic Radiology, Department of Translational Medicine, Lund University, Skåne University Hospital, SE-20502 Malmö, Sweden

Corresponding author
Ioannis Sechopoulos
Working address: Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Post 766, Nijmegen, The Netherlands
E-mail address: Ioannis.sechopoulos@radboudumc.nl
Telephone number: +31 24 366 80 89

List of Abbreviations

AI = Artificial intelligence

AUC = Area under the receiver operating characteristic curve

BI-RADS = Breast imaging reporting and data system

CAD = Computer-aided detection

CI = Confidence interval

DM = Digital mammography

MRMC = Multi-reader multi-case

PoM = Probability of malignancy

ROC = Receiver operating characteristic

SE = Standard error

Abstract

Background

Artificial intelligence (AI) systems performing at radiologist-like levels in the evaluation of digital mammography (DM) would improve breast cancer screening accuracy and efficiency. We aimed to compare the stand-alone performance of an AI system to that of radiologists in detecting breast cancer in DM.

Methods

Nine multi-reader multi-case study datasets previously used for different research purposes in seven countries were collected. Each dataset consisted of DM exams acquired with systems from four different vendors, multiple radiologists' assessments per exam, and ground truth verified by histopathological analysis or follow-up, yielding a total of 2,652 exams (653 malignant) and interpretations by 101 radiologists (28,296 independent interpretations). An AI system analyzed these exams yielding a level of suspicion of cancer present between 1 and 10. The detection performance between the radiologists and the AI system was compared using a non-inferiority null hypothesis at a margin of 0.05.

Results

The performance of the AI system was statistically non-inferior to that of the average of the 101 radiologists. The AI system had a 0.840 (95% CI = 0.820-0.860) area under the ROC curve (AUC) while the average of the radiologists was 0.814 (95% CI = 0.787-0.841) (difference 95% CI = (-0.003-0.055)). The AI system had an AUC higher than 61.4% of the radiologists.

Conclusions The evaluated AI system achieved a cancer detection accuracy comparable to an average breast radiologist in this retrospective setting. While promising, the performance and impact of such a system in a screening setting needs further investigation.

Keywords Mammography, artificial intelligence, breast cancer, computer detection systems, deep learning

Introduction

Breast cancer is the most common cancer in women, and despite important improvements in therapy, it is still a major cause for cancer-related mortality, accounting for approximately 500,000 annual deaths worldwide[1]. Population-based breast cancer screening programs using mammography are regarded as effective in reducing breast cancer-related mortality [2-5]. However, current screening programs are highly labor intensive due to the large number of women screened per detected cancer, and the use of double reading, especially in European screening programs, which also leads to additional economical costs. Moreover, despite this practice up to 25% of mammographically-visible cancers are still not detected at screening [6-9].

Considering the increasing scarcity of radiologists in some countries, including breast screening radiologists [10-12], alternative strategies to allow continuation of current screening programs are required. In addition, it is of paramount importance to prevent visible lesions in digital mammography (DM) being overlooked or misinterpreted.

Since the 1990s, computer-aided detection (CAD) systems have been developed to automatically detect and classify breast lesions in mammograms. The widespread implementation of DM for breast cancer imaging further spurred the development of automated detection techniques for breast cancer. Unfortunately, no studies to date have found that traditional CAD systems directly improve screening performance or cost-effectiveness, mainly because of a low specificity [13,14]. This has also precluded their use as a stand-alone reader for screening mammography.

However, the field of artificial intelligence (AI) is rapidly changing due to the success of novel algorithms based on deep learning convolutional neural networks. These approaches are very successful in automating cognitively difficult tasks; classic examples include self-driving cars and advanced speech recognition. In medical imaging, deep learning-based AI is also rapidly closing the gap between humans and computers [15,16]. It has been suggested that such algorithms could therefore have the potential to further improve the benefit-harm-ratio of breast cancer screening programs [17]. In recent years, several deep learning-based algorithms for automated analysis of mammograms have been developed, some of which have already shown very promising results when compared to radiologists, but in very limited and homogeneous scenarios [18,19].

Therefore, in this study, we compare, at a case level, the cancer detection performance of a commercially available AI system to that of 101 radiologists who scored nine different cohorts of DM examinations from four different manufacturers as part of reader studies previously performed for other purposes.

Methods

Artificial intelligence system

In this study we used an AI system for breast cancer detection in DM and digital breast tomosynthesis (Transpara 1.4.0, Screenpoint Medical BV, Nijmegen, The Netherlands). The system uses deep learning convolutional neural networks, feature classifiers and image analysis algorithms to detect calcifications[20,21] and soft tissue lesions[22-24] in two different modules. For each exam, on the basis of the individually-classified suspicious findings, the system provides a continuous score ranging between 1 and 10 representing the level of suspicion of cancer present (where 10 represents highly suspicious of malignancy present). This system can be applied to processed (i.e. "for presentation") DM images from multiple vendors and makes use of both the mediolateral oblique and cranio-caudal views of each breast. However, the AI system does not use information from prior mammograms (when available).

The AI system is trained, validated, and tested using a database containing over 9,000 mammograms with cancer (one third of which are presented as lesions with calcifications) and 180,000 mammograms without abnormalities. The mammograms originate from devices from four different vendors (Hologic; Siemens; General Electric, Waukesha, WI; Philips, Eindhoven, The Netherlands) and institutions across Europe, United States and Asia. The AI system is independently tested with exams never used for training or validation of the algorithms. The mammograms used in this study have never been used to train, validate or test the algorithms.

Digital Mammograms

We collected sets of DM examinations that were read by multiple radiologists during other unrelated, and previously completed, retrospective multi-reader multi-case (MRMC) observer studies [25-32]. In those studies, DM was compared to another modality (e.g. digital breast tomosynthesis) for breast cancer detection in cancer-enriched datasets. In total, nine distinct DM datasets were obtained from different institutions across Europe and the United States (**Table 1**). The review board at each institution waived local ethical approval and informed consent or directly approved the use of the anonymized patient data for retrospective research.

Each dataset consisted of three items: DM exams, the radiologists' scores of each DM exam, and their ground truth. DM exams were processed "for presentation" 2D images, two views per breast (CC and MLO) that could be unilateral or bilateral The corresponding radiologists' scores for each DM exam were in the form of forced Breast Imaging Reporting and Data System (BI-RADS®) scores (scale 1-5; 1 = negative, 2 = benign findings, 3 = probably benign, 4 = suspicious abnormality, 5 = highly suspicious of malignancy) and/or probability of malignancy (PoM) scores (scale 1-100). All interpretations involved single reading by individual radiologists, differing from standard practice in many screening programs, which use double reading plus consensus or arbitration. Finally, the ground truth was defined in terms of cancer present or absent, of each DM exam, confirmed by histopathology and/or at least one year of follow-up.

In all datasets, the radiologists individually scored each DM exam, without time constraint, and without access to other imaging techniques or any AI systems. There were differences across datasets (see **Table 1**) regarding study population and reading workflow.

Also, for some datasets, the radiologists had access to priors (not processed by this version of this AI system). In total, 28,296 independent exam interpretations of 2,652 cases were collected. Differences in numbers between the original study populations and the included populations are due to images and/or readings lost during data archiving at the original institutions (n = 13) as well as problems during processing with the AI system (n = 7, e.g., because the case contained implants).

Table 1 shows the distributions of the radiologists' experience with mammography for each dataset, which resembles the heterogeneous distribution seen in practice, as reported in the original publications. Readers from the United States were MQSA-qualified (Mammography Quality Standards Act), and included an approximately even mix of general and breast-specialized radiologists, while all the readers from Europe were specialized in breast imaging and were qualified according to the European guidelines for quality assurance in breast cancer screening [33]. For their studies, they were instructed to score simulating a screening practice.

Statistical analysis

The accuracy of the radiologists was compared to that of the AI system with a non-inferiority null-hypothesis based on differences in the area under the receiver operating characteristic curve (AUC). Only cases with malignant lesions were considered positive. Since this AI system had not been tested before, we did not assume a performance level pre-study, and hence did not calculate the power of this study. Instead, the study was performed with as much data as could be gathered to have the most robust conclusion possible.

Non-inferiority testing

Non-inferiority analysis [34-38] was used to compare the AI system to the radiologists. The non-inferiority margin was set at 0.05, since it was considered that differences below this margin are clinically unimportant. Non-inferiority was concluded if the AUC difference AI-radiologists was greater than 0 and the lower limit of the 95% confidence interval (CI) of the difference was greater than the negative value of the non-inferiority margin (-0.05).

Primary endpoint: Overall AUC performance of the AI system vs. 101 radiologists

We pooled the datasets listed in **Table 1** and compared the reader-averaged AUC against the AUC of the AI system. The public-domain iMRMC software (version 4.0.0, Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, MD) [36,37] was used, which can handle arbitrary (not fully-crossed) study designs, including the split-plot design resulting when pooling datasets as in this study [39,40]. The software expects multiple readers but can treat a single reader (the AI system) if the data is formatted properly. The iMRMC software can also handle the mixed scoring scales in the different datasets since the scores from different readers are never compared. If probability of malignancy was available, it was preferred over BI-RADS as it better samples the receiver operating characteristic (ROC) space and it is an ordinal scale [41]. For the AI system, its scoring exam-based scale (1-10) was used for ROC analysis. We created reader-averaged ROC curves by averaging the reader-specific non-parametric (trapezoidal) curves along lines perpendicular to the chance line [42]. This average is area-preserving; its AUC is equal to the reader-averaged non-parametric AUCs.

The analysis of the MRMC data, which yielded the empirical AUC values and their 95% CI, were computed following U-statistics to provide unbiased estimates of the variance components[36,43]. In this way, the total variance is decomposed into eight moments from first principles (similar to U-statistics), considering non-diseased cases separately from diseased cases so that the total variance can be easily generalized to new readers, new non-diseased cases, and new diseased cases.

Secondary endpoints: performance comparisons for each dataset

As secondary endpoints, the AUC and operating points were compared between the AI system and the average of radiologists for each dataset and against each individual radiologist. The reported 95% CI are not adjusted for testing multiple hypotheses, since the high amount of multiple comparisons (N=215) would make statistical testing impractical. Instead, this analysis is meant to be descriptive and to identify any possible outliers in the datasets.

Standard MRMC analysis of variance was used to compare the AUC between the AI system and the average of radiologists, based on the methods by Gallas et al. implemented in iMRMC [36,37]. Similarly, as with the split-plot analysis defined above, the AI system was defined as an independent second modality.

The sensitivity at the radiologists' specificity was compared between the radiologists and the AI system as determined by a screening scenario threshold (BI-RADS 3 or higher was considered positive, while in dataset C, radiologists directly indicated whether the case was recalled or not). There was no recall information for Dataset B, which involved 6

radiologists, (the original study did not ask radiologists for a recall decision) and therefore it was not included in this analysis. Consequently, sensitivity could therefore only be computed for 95 radiologists. The average sensitivity and specificity of the radiologists were computed with iMRMC using a single-modality analysis of variance with dichotomized scores as input. For the AI system, the operating point of the ROC that was closest to the average radiologist's specificity was then selected to dichotomize the results. Radiologists and AI system sensitivities were compared with iMRMC using a standard MRMC two-modality analysis of variance at the same specificity level.

Results

Overall AUC performance: AI system vs. 101 radiologists

The AUC of the AI system (0.840, 95% CI = 0.820-0.860) was statistically non-inferior to that of the 101 radiologists (0.814, 95% CI = 0.787-0.841). The AUC difference was 0.026 (95% CI = -0.003, 0.055), slightly higher for the AI system at the range of low- and mid-specificity. The average ROC curves are displayed in **Figure 1**.

The system had a higher AUC than 62/101 radiologists (61.4%, **Figure 2**) and higher sensitivity than 55/95 radiologists (57.9%, **Figure 3**), but its performance was always lower than that of the best radiologist **(Supplementary Table 1).**

Performance comparisons for each dataset

For each dataset, the AUC and sensitivity of the AI system was similar to that of the average of the radiologists, and no outliers were identified (**Supplementary Tables 1 and 2**). Absolute differences (AUC AI system – AUC average of radiologists) varied between -0.008 and +0.038 per dataset (**Supplementary Table 1**). The ROC curve of the AI system is plotted against the radiologists' ROC curves in **Supplementary Figure 1**.

The average operating point of the radiologists was different across datasets, with specificities ranging from 0.49 to 0.79, and sensitivities between 0.76 and 0.84 (see **Supplementary Table 2 and Supplementary Figure 1**). At the average specificity of the radiologists, the AI system had a higher sensitivity in 5 out of 8 datasets (1.0-8.0%), and lower in 3 datasets (1.0-2.0%).

Discussion

Our results clearly show that recent advances in AI algorithms have narrowed the gap between computers and human experts in detecting breast cancer in digital mammograms. Nevertheless, the performance of AI was consistently lower than the best radiologists in all datasets. The large and heterogeneous population of cases used in this study shows that our findings might hold true across different lesion types, mammography systems and country-specific practices.

Across the collected data, differences were seen in the performance of the readers. As expected, readings in the United States had a lower average specificity than those in Europe, where screening recall rates are lower [44]. For Dataset A, even though performed in Europe, the average specificity is similar to North-American readings. Perhaps this is

explained by the dataset being mostly composed of breasts with high density, which might have made radiologists modify their operating points. The wide range in average AUC values (0.769 – 0.907) across datasets shows that the difficulty of the populations varied substantially, due to, for instance, inclusion of specific lesion types, different proportions of enrichments, or availability of prior exams and/or exams of the contralateral breast. It should be noted that the AUC values for the radiologists were lower than those reported in US clinical practice by the Breast Cancer Surveillance Consortium, which are above 0.90 [45]. This is likely because the datasets used in this study were highly enriched with cancers and false positive exams, resulting in a case set which is substantially more challenging than a screening mammography set.

For the AI system, the performance was very close to the average of radiologists in all datasets. Interestingly, this also held in all datasets (Dataset B, C, D) where the AI system had the disadvantage of not considering information from the prior mammograms, whereas the radiologists had access to available prior images. The reader-averaged ROC curve of the 101 radiologists was almost identical to that of the AI system at high specificity, while the AI system showed slightly higher AUC at mid and low specificity. Since this data was enriched with cancer and benign lesions, the screening recall operating point of radiologists lied at the mid-range in specificity. At this fixed recall specificity, the AI system achieved higher sensitivity than a majority of the radiologists.

However, given the fact that this database was not prospectively defined for this study, caution should be taken in interpreting the results. In particular, although most exams in the original studies are from screening, and all radiologists were instructed to

score simulating a screening practice, the main limitation of this study is that it was based on retrospective reader studies of enriched case sets. Therefore, the human performance was affected by a "laboratory effect" that reflects the reading of enriched datasets [46,47]. Since the main application of such an AI system would be a screening setting, the stand-alone performance of the AI system on actual screening data should be studied, including the distribution of lesions seen in screening, and comparing it to the radiologists' performance during actual screening interpretation. Collecting such a high number of cancer cases and prospective readings from a similarly large number of radiologists in an actual screening scenario, would be notably challenging, however, requiring the collaboration of a very large number of centers.

Even if the AI system performed comparably to the human radiologists, there is still room for improvement. There is no *a priori* reason why the AI system should not be performing, at least, as the best radiologist. In our study, the AI system had an AUC lower than the best radiologist in every dataset. This could be explained by the fact that radiologists interpret more information (e.g. comparisons with prior exams and contralateral breasts) than this version of this AI system. An ideal AI system should be able to perform up to the limitations of the imaging modality itself; in other words, be only incapable of detecting mammographically occult cancers, while minimizing false positive findings. Determining the trade-off between cancer detection and assessment of false positive findings would then be the only human choice involved. However, to achieve a higher-than-human performance, the training of the AI systems might need to not be based on truth as established by humans.

Future work, not assessed in our study due to lack of information from the original studies, is to analyze the AI system performance per lesion type, tumor characteristics, or lesion location. For instance, evaluation of the sensitivity as a function of false positive findings, taking localization into account (i.e., using free receiver operating characteristic analysis) should could be of interest, especially in order to verify the potential of using such an AI system as a reader aid rather than as a stand-alone reader. Moreover, although most cases were collected from screening examinations, a limitation is that we cannot know exactly how representative of an actual screening population our dataset is, in terms of tumor size and types, since these characteristics were not reported in the original study publications. Similarly, it is unknown whether the better performing radiologists were the radiologists with the most experience, as the original studies did not report the individual experience of each radiologist. Consequently, we cannot assess whether the AI system performs better or worse than radiologists as a function of the experience of the latter. However, the heterogeneity of experience seen in our data is representative of that seen in screening practice. Consequently, we can conclude that the AI system is as good as an average screening radiologist.

Artificial intelligence that functions at the level of an expert radiologist for breast cancer detection in DM images might herald a change in the breast healthcare workflow, whether in a screening or in a clinical setting. Yet we still need to determine the optimal integration of such a system in the breast care pathway, prior to assessing the final impact that this type of AI technology can have on patient care.

In a population-based screening setting, the possibilities of workflow enhancement via implementation of an AI system are ample. One of the biggest potential benefits lies in the possibility of using such a system in countries where there is a lack of experienced breast radiologists, which might, for instance, impede the development, expansion, or continuation of screening programs. In these situations, AI could be used as an independent stand-alone first or second reader [48].

In parallel, it could also be used as an interactive decision support tool [27], pointing out potential lesions, preventing overlook and interpretation errors that are relatively common in the reading of DM [6-9]. However, for this aspect, the impact of automation bias in decision making should be addressed. Furthermore, it is well known that the very low prevalence of breast cancer in the screening population reduces the performance of radiologists, increasing the risk of false negatives [47,49]. An AI system tuned to achieve high sensitivity could be used to automatically discard a significant amount of DM exams which are most likely normal, reducing the workload and resulting in a case set with a higher prevalence of cancer for radiologists to read. The higher sensitivity of the AI system at low specificity found in this study points to the feasibility of this scenario. However, the drawbacks of introducing AI, especially as stand-alone readers, have to be studied. Regulations to define the medicolegal consequences when AI fails would have to be established. Equally, trade-offs between patient outcome and cost-effectiveness have to be carefully addressed.

In conclusion, the tested AI system based on deep learning algorithms has similar performance as an average radiologist for detecting breast cancer in mammography. These

results were consistently observed across a large heterogeneous multi-center multi-vendor cancer-enriched cohort of mammograms. While promising, the performance and the fashion of implementation of such an AI system in a screening setting remains to be further investigated.

Funding

This work did not receive any funding.

Relevant conflict of interest: TM reports to be an employee of Siemens Healthineers. AGM and TT report to be employees of ScreenPoint Medical BV. SZ, PC, TH, KL, RM and IS report to have received research funding, unrelated to this work, from Siemens Healthineers. During the period of the study, MC, MB, ARR, IA, MW report no conflict of interest.

Acknowledgements: The authors would like to thank Dr. Brandon Gallas, Dr. Weijie Chen, and mr. Qi Gong (Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, MD, USA) for help implementing the statistical methods of the study with their iMRMC software (https://github.com/DIDSR/iMRMC). We would also like to thank all the radiologists involved in the reader studies whose results were used in this work; Britta Stenson (Philips Healthcare, Stockholm, Sweden) for the help gathering data; and ScreenPoint Medical for providing their software and technical support for this research.

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer.* 2015;136(5).
2. Broeders M, Moss S, Nyström L, et al. The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies. *Journal of medical screening.* 2012;19(1_suppl):14-25.
3. Lauby-Secretan B, Scoccianti C, Loomis D, et al. International Agency for Research on Cancer Handbook Working Group et al (2015) Breast Cancer Screening–Viewpoint of the IARC Working Group. *N Engl J Med.*372(24):2353-2358.
4. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer.* 2013;108(11):2205-2240.
5. Smith RA, Andrews KS, Brooks D, et al. Cancer screening in the United States, 2017: a review of current American Cancer Society guidelines and current issues in cancer screening. *CA: a cancer journal for clinicians.* 2017;67(2):100-121.
6. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology.* 1992;184(3):613-617.
7. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics.* 2003;23(4):881-895.
8. Weber RJ, van Bommel RM, Louwman MW, et al. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. *Breast cancer research and treatment.* 2016;158(3):471-483.

9. Broeders M, Onland-Moret N, Rijken H, Hendriks J, Verbeek A, Holland R. Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. *European Journal of Cancer.* 2003;39(12):1770-1775.

10. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online).* 2017;359.

11. National Health Institutes England, Public Health England, British Society of Breast Radiology, Royal College of Radiologists. The breast imaging and diagnostic workforce in the United Kingdom. 2017; https://www.rcr.ac.uk/publication/breast-imaging-and-diagnostic-workforce-united-kingdom.

12. Wing P, Langelier MH. Workforce shortages in breast imaging: impact on mammography utilization. *American Journal of Roentgenology.* 2009;192(2):370-378.

13. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med.* 2007;356(14):1399-1409.

14. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern Med.* 2015;175(11):1828-1837.

15. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60-88.

16. Bejnordi BE, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama.* 2017;318(22):2199-2210.

17. Trister AD, Buist DSM, Lee CI. Will Machine Learning Tip the Balance in Breast Cancer Screening? *JAMA Oncol.* 2017.

18. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Invest Radiol.* 2017;52(7):434-440.

19. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal.* 2017;35:303-312.

20. Mordang J-J, Janssen T, Bria A, Kooi T, Gubern-Mérida A, Karssemeijer N. Automatic Microcalcification Detection in Multi-vendor Mammography Using Convolutional Neural Networks. Paper presented at: International Workshop on Digital Mammography (IWDM)2016; Malmo, Sweden.

21. Bria A, Karssemeijer N, Tortorella F. Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications. *Med Image Anal.* 2014;18(2):241-252.

22. Hupse R, Karssemeijer N. Use of normal tissue context in computer-aided detection of masses in mammograms. *IEEE Trans Med Imaging.* 2009;28(12):2033-2041.

23. Karssemeijer N. Automated classification of parenchymal patterns in mammograms. *Phys Med Biol.* 1998;43(2):365-378.

24. Karssemeijer N, Te Brake GM. Detection of stellate distortions in mammograms. *IEEE Trans Med Imaging.* 1996;15(5):611-619.

25. Wallis MG, Moa E, Zanca F, Leifland K, Danielsson M. Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution X-ray imaging observer study. *Radiology.* 2012;262(3):788-796.

26. Visser R, Veldkamp WJ, Beijerinck D, et al. Increase in perceived case suspiciousness due to local contrast optimisation in digital screening mammography. *Eur. Radiol.* 2012;22(4):908-914.

27. Hupse R, Samulski M, Lobbes MB, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. *Radiology.* 2013;266(1):123-129.

28. Gennaro G, Hendrick RE, Ruppel P, et al. Performance comparison of single-view digital breast tomosynthesis plus single-view digital mammography with two-view digital mammography. *Eur Radiol.* 2013;23(3):664-672.

29. Siemens Medical Solutions USA Inc. FDA Application: Mammomat Inspiration with Digital Breast Tomosynthesis. 2015; https://www.accessdata.fda.gov/cdrh_docs/pdf14/P140011b.pdf. Accessed March 3, 2018.

30. Garayoa J, Chevalier M, Castillo M, et al. Diagnostic value of the stand-alone synthetic image in digital breast tomosynthesis examinations. *Eur Radiol.* 2018;28(2):565-572.

31. Rodriguez-Ruiz A, Gubern-Merida A, Imhof-Tas M, et al. One-view digital breast tomosynthesis as a stand-alone modality for breast cancer detection: do we need more? *Eur Radiol.* 2017.

32. Clauser P, Baltzer P, Kapetas P, et al. Synthetic 2D mammography can replace digital mammography as an adjunct to wide-angle digital breast tomosynthesis. *Investigative radiology.* 2018;Accepted.

33. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. —summary document. *Annals of Oncology.* 2008;19(4):614-622.

34. Blackwelder WC. “Proving the null hypothesis” in clinical trials. *Controlled clinical trials.* 1982;3(4):345-353.

35. Chen W, Petrick NA, Sahiner B. Hypothesis testing in noninferiority and equivalence MRMC ROC studies. *Academic radiology.* 2012;19(9):1158-1165.

36. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators. *Communications in Statistics—Theory and Methods.* 2009;38(15):2586-2603.

37. Gallas B. iMRMC v4.0: Application for Analyzing and Sizing MRMC Reader Studies. 2017; https://github.com/DIDSR/iMRMC/releases, https://cran.r-project.org/web/packages/iMRMC/index.html.

38. Gennaro G. The "perfect" reader study. *European journal of radiology.* 2018;In press.

39. Chen W, Gong Q, Gallas BD. Efficiency gain of paired split-plot designs in MRMC ROC studies. Paper presented at: Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment2018.

40. Gallas BD, Brown DG. Reader studies for validation of CAD systems. *Neural Netw.* 2008;21(2-3):387-397.

41. Jiang Y, Metz CE. BI-RADS Data Should Not Be Used to Estimate ROC Curves. *Radiology.* 2010;256(1):29-31.

42. Chen W, Samuelson FW. The average receiver operating characteristic curve in multireader multicase imaging studies. *The British journal of radiology.* 2014;87(1040):20140016.

43. Gallas BD, Hillis SL. Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances. *Journal of Medical Imaging.* 2014;1(3):031006.

44. Skaane P. Breast cancer screening with digital breast tomosynthesis. *Breast Cancer.* 2017;24(1):32-41.

45. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *Journal of the National Cancer Institute.* 2004;96(24):1840-1850.

46. Gur D, Bandos AI, Cohen CS, et al. The “laboratory” effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology.* 2008;249(1):47-53.

47. Evans KK, Birdwell RL, Wolfe JM. If you don’t find it often, you often don’t find it: Why some cancers are missed in breast cancer screening. *PLoS One.* 2013;8(5):e64366.

48. Gilbert FJ, Astley SM, Gillan MG, et al. Single reading with computer-aided detection for screening mammography. *New England Journal of Medicine.* 2008;359(16):1675-1684.

49. Huynh PT, Jarolimek AM, Daye S. The false-negative mammogram. *Radiographics.* 1998;18(5):1137-1154; quiz 1243-1134.

Table

Table 1. Details of each dataset collected for this study.

Dataset	Reference	Reading Country	Vendor(s)	Case set population	Exam type	Total no. of exams	Exam result, No.			No. of Radiologists	Radiologists Experience, y	Score scale
							Cancer	Benign lesions	Normal			
A [25]	Wallis et al, 2012 [25]	Sweden, UK	GE Sectra	40-80 (avg.=56) years old Screening (n=86) + Clinical (n=43) Only BI-RADS density >2	Bilateral no priors	129	40	23	66	14	3-25 (avgerage= 10)	BI-RADS
B [26]	Visser et al, 2012,[26]	Netherlands	GE	51-86 (avg.=60) years old Screening	Bilateral + priors	263	43	110	110	6	1-34	PoM*
C [27]	Hupse et al, 2013[27]	Netherlands	Hologic	50-74 years old Screening	Bilateral + priors	199	79	20	100	9	1-24 (average = 14)	PoM†
D [28]	Gennaro et al, 2013[28]	Italy	GE	>40 years old Clinical	Unilateral + priors	469	68	200	201	6	5-30	BI-RADS
E1 [29]	Siemens Medical Solutions, 2015[29]	US	GE Siemens Hologic	>40 years old Screening + Clinical	Bilateral no priors	298‡	49	84	165	22	> 5	PoM BI-RADS
E2 [29]	Siemens Medical Solutions, 2015[29]	US	GE Siemens Hologic	>40 years old Screening + Clinical	Bilateral no priors	326‡	104	79	143	31	> 5	PoM BI-RADS

F [30]	Garayoa et al, 2018 [30]	Spain	Hologic	34-92 (avg.=55) years old Screening + Clinical	Unilateral no priors	585	113	160	313	3	10-20	BI-RADS
G [31]	Rodriguez-Ruiz et al, 2018 [31]	Netherlands Sweden	Siemens	30-88 (avg.=52) years old Screening (n=60) + Clinical (n=121)	Unilateral no priors	179	75	49	55	6	3-44 (average = 22)	PoM BI-RADS
H [32]	Clauser et al, 2018, [32]	Austria	Siemens	36-84 (avg.=56) years old Screening + Clinical	Bilateral no priors	204	82	43	80	4	> 5	BI-RADS
TOTAL	-	7 countries	4 vendors	-	-	2652	653 (24.6%)	768 (29.0%)	1233 (46.4%)	101	-	-

* No BI-RADS scores were used in this study, and radiologists were not asked to decide on recall/no recall. PoM = Probability of malignancy (1-100); avg. = average. BI-RADS: Breast Imaging Reporting and Data System scores (1-5).

DM manufacturers listed: Sectra Mamea, Solna, Sweden; Siemens Healthineers, Forcheim, Germany; Hologic Inc, Bedford, MA, USA; General Electric Healthcare, Waukesha, WI, USA

† No BI-RADS scores were used in this study, but radiologists were asked to decide on recall/no recall

‡ The cases from these two datasets overlap and come from a unique population of 425 DM exams (107 malignant, 102 benign, 216 normal). The radiologists are different.

Figures

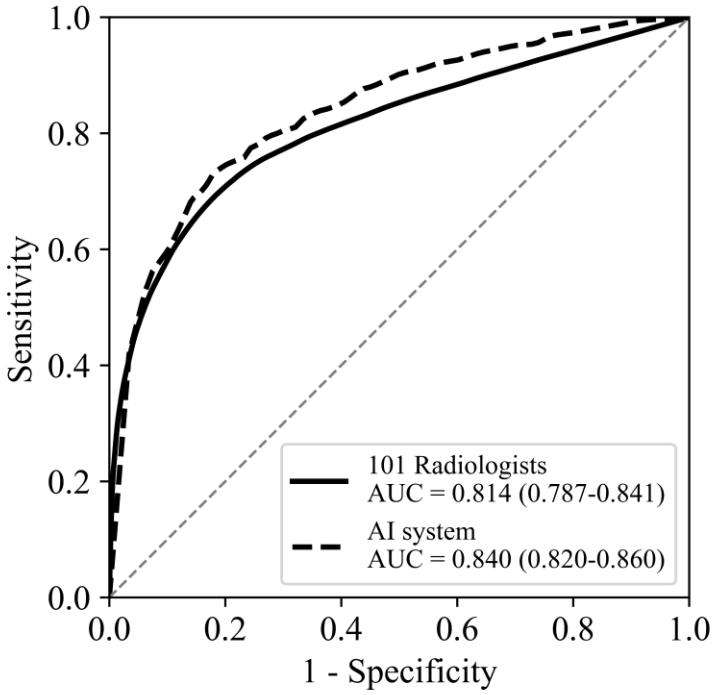

Figure 1. Receiver operating characteristic (ROC) curve comparison between the reader-averaged radiologists and the artificial intelligence (AI) system in terms of area under the curve (AUC). Parentheses show the 95% confidence interval of the AUC.

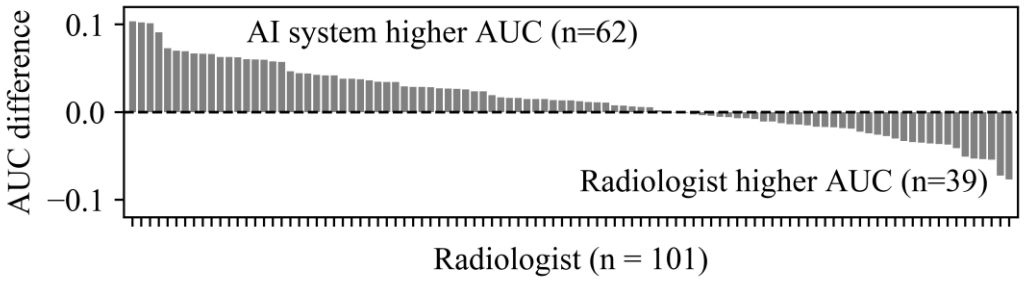

Figure 2. Differences in area under the receiver operating characteristic curve (AUC) between the artificial intelligence (AI) system and each radiologist.

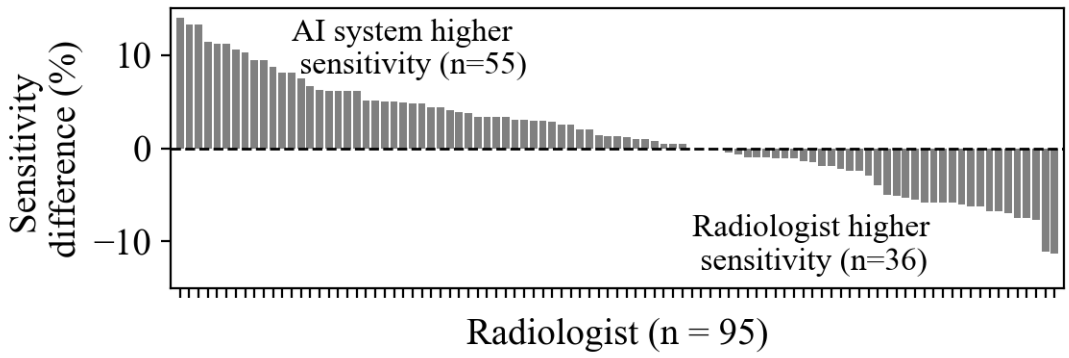

Figure 3. Differences (%) in sensitivity between the artificial intelligence (AI) system and each radiologist, at the specificity of each radiologist considering BI-RADS ≥3 as positive recall. BI-RADS = Breast Imaging Reporting and Data System.